基于语义融合与模型蒸馏的农业实体识别

李亮德[1,2]，王秀娟[1,3]，康孟珍[1,2*]，华　净[1,4]，樊梦涵[1,2]

（1. 中国科学院自动化研究所 复杂系统管理与控制国家重点实验室，北京 100190；2. 中国科学院大学 人工智能学院，北京 100049；3. 北京智能化技术与系统工程技术研究中心，北京 100190；4. 青岛中科慧农科技有限公司，山东青岛 266000）

摘　要：当前农业实体识别标注数据稀缺，部分公开的农业实体识别模型依赖手工特征，实体识别精度低。虽然有的农业实体识别模型基于深度学习方法，实体识别效果有所提高，但是存在模型推理延迟高、参数量大等问题。本研究提出了一种基于知识蒸馏的农业实体识别方法。首先，利用互联网的海量农业数据构建农业知识图谱，在此基础上通过远程监督得到弱标注语料。其次，针对实体识别的特点，提出基于注意力的BERT层融合模型（BERT-ALA），融合不同层次的语义特征；结合双向长短期记忆网络（BiLSTM）和条件随机场CRF，得到BERT-ALA+BiLSTM+CRF模型作为教师模型。最后，用BiLSTM+CRF模型作为学生模型蒸馏教师模型，保证模型预测耗时和参数量符合线上服务要求。在本研究构建的农业实体识别数据集以及两个公开数据集上进行实验，结果显示，BERT-ALA+BiLSTM+CRF模型的*macro-F*1相对于基线模型BERT+ BiLSTM+CRF平均提高1%。蒸馏得到的学生模型BiLSTM+CRF的*macro-F*1相对于原始数据训练的模型平均提高3.3%，预测耗时降低了33%，存储空间降低98%。试验结果验证了基于注意力机制的BERT层融合模型以及知识蒸馏在农业实体识别方面具有有效性。

关键词：远程监督；农业知识图谱；农业问答系统；实体识别；知识蒸馏；深度学习；BERT；双向长短期记忆网络

中图分类号：TP391　　**文献标志码：**A　　**文章编号：**202012-SA001

引用格式：李亮德, 王秀娟, 康孟珍, 华净, 樊梦涵. 基于语义融合与模型蒸馏的农业实体识别[J]. 智慧农业(中英文), 2021, 3 (1): 118-128.

LI Liangde, WANG Xiujuan, KANG Mengzhen, HUA Jing, FAN Menghan. Agricultural named entity recognition based on semantic aggregation and model distillation[J]. Smart Agriculture, 2021, 3 (1): 118-128. (in Chinese with English abstract)

1　引　言

随着农业互联网的发展与农业从业人员的新老更替，需要进行农业知识的快速传播和应用，以解决农业技术人员不足的问题。目前，互联网上的农业知识技术问答主要由人工专家来完成，这样不仅效率低，而且受技术专家资源稀缺的限制。如果计算机能够理解用户输入的农业问题，

收稿日期：2020-12-17　修订日期：2021-02-09

基金项目：中国科学院战略性先导科技专项（A类）（XDA20030102）；国家自然科学基金面上项目（62076239）；中国科学院与泰国科技发展署合作研究资助项目（GJHZ2076）

作者简介：李亮德（1996—），男，硕士研究生，研究方向为面向农业的自然语言处理。E-mail：liliangde2018@ia.ac.cn。

*通讯作者：康孟珍（1975—），女，博士，副研究员，研究方向为计算植物和智慧农业。电话：010-82544502。E-mail：mengzhen.kang@ia.ac.cn。

通过建立农业知识图谱来进行智能回答，将大大提高农业知识问答效率。

农业智能问答系统包括信息抽取[1]、知识图谱构建、问句理解和基于知识库的问答四个环节。信息抽取用于理解问题并基于农业知识图谱回答问题，对于农业智能问答系统至关重要。命名实体识别[2]是指识别出文本中的实体指称项及其类别，是自然语言处理中一项基础任务。基于农业实体识别可抽取文本中关键信息，构建农业知识图谱，实现农业知识结构化，进而基于知识图谱进行农业知识问答。互联网上储存着大量的非结构化农业文本，如何将这些杂乱无章的农业文本转化成结构化的农业知识，构建农业知识图谱，是实现农业智能问答系统的重要环节。

农业知识数据特别是标注数据难以获取，有关农业知识图谱构建以及信息抽取的研究相对较少。已有研究的农业实体识别方案往往需要大量的训练数据训练，因此应用这些方案时，需要人工标注实体识别数据，成本很高。使用的模型也存在需要手工提取特征、实体识别效果不佳等问题；或者没有考虑实际线上对预测耗时、模型大小的要求，停留在实验验证阶段。李贯峰和张鹏[3]使用词典来实现实体识别，构建了基于农业本体的web知识抽取模型，因为web知识库难以覆盖所有的农业实体，因此存在着召回率低的缺点。王春雨和王芳[4]用条件随机场[5]来进行命名实体识别。但这种方法需要手工构造特征且模型容量低，难以完成复杂的实体识别任务。印度的Malarkodi等[6]应用了条件随机场模型，输入一些句法词汇特征，同样存在依赖手工构造特征的问题。刘晓俊[7]使用了基于密集连接的双向长短记忆网络（Dense Connected Bi-directional Long Short-Term Memory，DC-LSTM）+（Conditional Random Field，CRF）架构进行面向农业领域的命名实体识别研究。由于这是一种多层的密集连接的结构，推理耗时长、模型参数量多，难以在线上进行实际使用。Biswas等[8]利用WordNet[9]进行农业实体识别，该方法本质上与词典匹配差别不大，但是利用WordNet词的相关性，扩充了词典。

目前，无论是基于条件随机场等传统方法，还是基于深度学习[10]的实体识别模型，都是数据驱动的，需要海量的标注数据作为支撑。在农业领域缺乏大量现成的标注数据情况下，直接套用通用领域的实体识别方案难以奏效。因此，本研究提出了一种基于远程监督[11]的农业领域数据标注方案，以解决农业实体识别标注数据稀缺的问题。

远程监督的思想由Mintz于第47届计算语言协会年会上（Association for Computational Linguistics）上首次提出，通过将知识库与文本对齐来自动构建大量训练数据，减少模型对人工标注数据的依赖，增强模型跨领域适应能力，被大量运用在关系抽取领域[12]。远程监督提出的动机是解决关系抽取标注数据难以获取的问题，而农业实体识别数据存在标注数据难以获取的问题，因此本文将远程监督的思想迁移到实体识别领域。通用领域具有一词多义性质，在通用领域给远程监督带来很大的噪声。但是，在农业等专有领域，虽然存在漏标注的情况，但是词的语义固定，整体上噪声比较小，因此远程监督是可行的方案，可以很好地规避农业领域缺乏标注数据的问题。

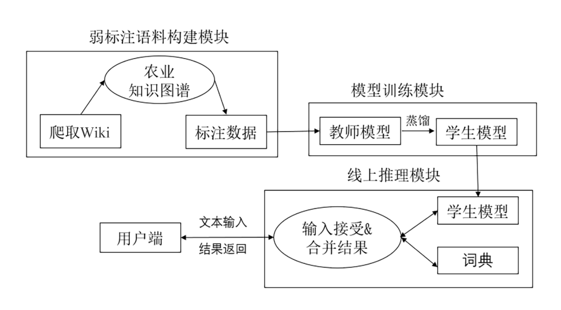
本研究采用目前在自然语言处理领域流行的大规模预训练模型基于转换器的双向编码表征（Bidirectional Encoder Representations from Transformers，BERT）[13]，一方面，预训练模型在海量互联网数据上的训练，模型容量大，能够拟合复杂的实体识别任务；另一方面，农业实体识别标注数据比较稀缺，而预训练模型基于大规模语料训练的，包含了很多基础的语言知识，在大规模预训练模型的基础上进行微调，农业实体识别模型也包含了这部分基础的语言知识。此外，本研究还结合农业实体识别的特点，提出了基于注意力的层融合机制（Attention-Based Layer Aggregation）对BERT做出改进。

在线问答系统需要时间和空间复杂度低的模型。前面提出了基于BERT的模型，但是BERT因为参数量大导致推理耗时高，很难满足实时推理需求。模型蒸馏[14]是将训练好的复杂模型推广“知识”能力迁移到一个结构更为简单的网络中，或者通过简单的网络去学习复杂模型中的“知识”。其中，训练好的复杂模型称为教师模型，而学习的简单模型称为学生模型。本研究考虑到模型上线对于预测耗时和模型大小的要求，用BiLSTM + CRF[15]作为学生模型，蒸馏前面得到的基于BERT的系列模型。

2　研究方法

2.1 整体架构

本研究提出的农业实体识别架构主要包括了弱标注[16]语料构建模块、模型训练模块以及线上推理模块（图1）。



注：Wiki为互联网上的多人协作的写作系统

图1　农业实体识别系统架构图

Fig.1 Architecture of agriculture named entity recognition

其中，弱标注语料构建模块采用了远程监督的思想，分为两个阶段：一是农业知识图谱构建阶段，爬取互联网的农业资源，过滤得到农业实体，构建农业知识图谱；二是数据弱标注阶段，通过前向最大匹配标注出文本里面的农业实体，用于模型训练。其中，模型训练模块又包含了两个阶段：一是教师模型训练阶段，用弱标注数据去训练本文提出的教师模型；二是模型蒸馏阶段[14]，用参数量少的模型作为学生模型蒸馏教师模型。线上推理模块接受用户端发送的文本，合并词典、学生模型的结果，返回给用户端。

2.2 数据来源

目前农业领域缺乏开源的中文农业知识图谱和农业实体识别语料。互动百科、百度百科都是开源的中文百科网站，包含了大量农业方面的实体和知识，很多农业网站上相关的农业知识也与百科网站上相同，不同的百科网站里面农业方面的知识类似。考虑到互动百科比其他百科类网站以及开源的农业信息网站更容易爬取，本研究选择爬取互动百科数据，建立农业知识图谱，用于构建农业知识图谱以及标注实体识别训练语料。将互动百科数据库下的农业实体对应的文档进行分句，得到农业实体识别语料。

2.3 基于远程监督的农业命名实体识别语料标注

将远程监督思想用在实体识别领域，是假设一个句子中的某个词与知识图谱里面的某个实体对应的名称或者别称相同，那么这个词就对应知识图谱里面的实体。远程监督的思想存在两方面的问题：首先，对于一词多义的实体会存在标注错误，例如把Apple手机的“苹果”对应到水果的“苹果”，但是一词多义在农业等专业领域的文本里面是可以忽略的；其次，对于不在农业知识图谱里面的实体，存在漏标注。通过远程监督方法对文本进行弱标注可以分为两个阶段：一是爬取互联网上多人协作的写作系统（Wiki）建立农业知识图谱，通过对Wiki本体的标签信息应用规则匹配，推断出实体的类型，过滤得到实体类型为作物、病害、农药等的实体，构建农业知识图谱；二是对语料进行弱标注，将农业知识图谱的实体用前缀树[17]保存起来，作为词典，对文本中的句子进行前向最大匹配，从而得到实体弱标注的结果。例如句子“怎样进行番茄分苗”通过前向最大匹配，就可以得到番茄两字对应农业知识图谱里面“番茄”这个实体，番茄实体的类别是作物（crop）。进而生成标签O（“怎”）O（“样”）O（“进”）O

（“行”）B_crop（“番”）I_crop（“茄”）O（“分”）O（“苗”）。其中，O（other）表示非实体，B（begin）表示实体开始位置，I（interior）表示实体内部以及结束位置，crop 表示实体类型为作物类型。B_crop I_crop 表示类型为作物的实体，分别对应于实体的开始和结束的位置，在句子中为第 4 和第 5 个词（“番茄”）。

2.4 教师模型

深度学习模型+条件随机场[15,18,19]是命名实体识别领域的主流模型[15]。深度学习模型指具有 BiLSTM[20]和 BERT[13]一类的模型，用于提取文本的语义特征，得到词到每个实体类别的概率；条件随机场用于计算各个实体类别的转移概率，结合生成概率和转移概率，进行 end2end 的训练。

2.4.1 BERT 模型

BERT 模型是谷歌 AI 团队于 2018 年发布的预训练模型，在 11 种不同自然语言处理验证任务中创造了最佳成绩。简单来说，BERT 在大量文本语料上使用自监督的方式训练了一个通用的语言理解模型，然后在这个模型上设置轻量级的下游任务接口去执行特定的自然语言处理任务。BERT 模型结构如图 2 所示。

BERT 模型主要包含三个部分：输入层、多转换器（transformer encoder），以及输出层。输入层由词嵌入（token-embedding）、位置嵌入（position-embedding）和段嵌入（segment-embedding）组成。词嵌入是将文本分为词，将词转化为向量；位置嵌入是将词的位置信息编码为特征向量，从而让模型获取到词的位置信息；段嵌入用于区分模型输入的两个句子。Transformer encoder[21]通过自注意力机制（self attention），实现词与词的相互交互，获得句子的语义表征。输出层在句子的语义表征基础上，根据下游任务来定具体的结构。BERT 训练分为预训练阶段和微调阶段两个阶段。在预训练阶段采用自监督的训练，主要任务是 Masked Language Model，也即随机掩盖句子里面的某些词，预测这些词，这个过程无需标注语料，可以直接通过互联网上的海量文本获取；在微调阶段，针对特定任务，设置不同的输出层和目标函数，利用少量标注数据进一步更新模型参数，即可完成针对特定领域的模型训练。

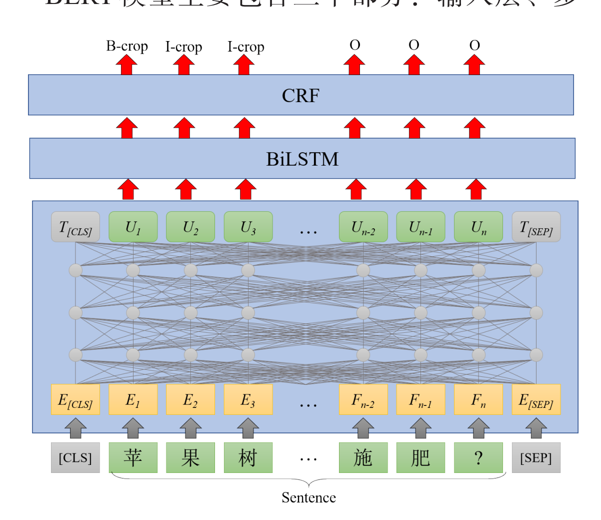


图 2　BERT 架构图

Fig. 2　Architecture of BERT

2.4.2 长短记忆网络

长短记忆（Long Short-Term Memory，LSTM）网络[20]用门机制去改善循环神经网络（Recurrent Neural Network，RNN）的梯度消失问题，双向循环神经网络（Bi-directional Long Short-Term Memory，BiLSTM）由两个单向的 LSTM 网络构成，两个网络中一个随时间正向传播，另一个随时间逆向传播。对于文本序列而言，BiLSTM 能有效的捕获上下文信息，在实体识别等序列标注任务上有效。

2.4.3 条件随机场

条件随机场（Conditional Random Field，CRF）模型[5]是一种概率无向图模型，可以解决序列标注任务。给定观察序列 X 的条件下求 Y，Y 隐状态序列的概率为 P（Y|X）。在命名实体识别上使用的 CRF 主要是 CRF 线性链，建模的数学公式下所示。

$$P(y|x)=\frac{1}{Z(x)}\exp(\sum_{k=1}^{K}w_k f_k(y,x)) \tag{1}$$

$$Z(x)=\sum_{y}\exp\sum_{k=1}^{K}w_k f_k(y,x) \quad (2)$$

其中，f_k是特征函数；w_k是特征函数的权重；$Z(x)$是归一化因子。模型在预测的时候使用维特比算法，这是一种动态规划算法，在给定观察序列X和参数的条件下，求出最大的标记序列Y的概率。

2.4.4 深度学习模型+条件随机场

深度学习模型本质上是把深度模型视为文本特征提取器，得到文本特征后，通过全连接层后得到词到实体类别的得分，记为P，输入到CRF层中。CRF层包含一个转移矩阵$\boldsymbol{A}$，表示两个标签的转移得分。模型对句子x标签等于y打分，打分经过softmax后得到概率，表达式如下所示。

$$score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1},y_i} \quad (3)$$

$$P(y|x)=\frac{\exp(score(x,y))}{\sum_{y'}\exp(score(x,y'))} \quad (4)$$

可以看出，整个句子的打分等于各个位置的打分之和，每个位置的打分由深度学习模型的输出P以及转移得分A决定。模型训练时，最大化对数损失函数即可。

深度模型可以是BERT、BiLSTM、空洞卷积神经网络（Iterated Dilated Convolutional Neural Network，IDCNN）等。目前在实体识别领域用的最多的是BERT和BiLSTM。BiLSTM+CRF 2016年由Dong等[22]提出，用于通用领域命名实体识别；BERT+CRF由Souza等[19]提出，用于葡萄牙语的命名实体识别；但是BERT的transformer的自注意力机制会破坏BERT的相对位置信息[23]。为解决BERT相对位置信息抽取能力不足的问题，一种方法是用BERT+BiLSTM[22]作为深度模型，BERT+BiLSTM+CRF由Jiang等[24]提出，用于通用领域的命名实体识别。BERT起到提供动态词向量的作用，BiLSTM用于建模相对位置信息。因此，本研究设置了三种基线模型BiLSTM + CRF[22]、BERT + CRF[19]和BERT + BiLSTM + CRF[24]来进行农业实体识别实验，进而选择出试验效果较好的模型作为教师模型，蒸馏轻量化的学生模型。这三种基线模型在其他领域都被验证有效。

2.4.5 基于注意力的BERT层融合模型机制

实体识别任务对于底层的语法、语义特征需求比较大，对于上层语义特征的需求反而没有那么强烈。BERT是一个多层transformer[21]的特征提取器，BERT-base模型一共包含了12层。多层transformer一方面减慢模型的推理速度。另一方面，Jawahar等[25]在ACL 2019发表的论文指出，BERT的低层网络学习到了短语级别的信息表征，BERT的中层网络学习到了丰富的语言学特征，而BERT的高层网络则学习到了丰富的语义信息特征。对于通用领域的实体识别而言，模型专注于顶层语义特征而忽视了实体识别任务亟需的底层特征。对于垂直领域，如农业的实体识别而言，判别实体的边界比判别实体的类别更难，因为垂直领域实体含义相对通用领域的判别容易一些。因此底层特征包含的短语级别的信息表征对于判别实体边界更重要，仅仅考虑顶层的高层语义信息显然不合理。另外一方面，本研究远程监督得到的标注数据的量有限，直接取高层的信息容易导致过拟合。因此，本研究提出一种基于注意力的BERT层融合机制。BERT模型包含多层transformer encoder，不同大小的BERT模型transformer encoder层数不同，一般有12、24、48三种，将BERT的层数记为L，做基于注意力机制的层融合，其中α和γ都是可训练的参数，如公式（5）和公式（6）所示。

$$h=\gamma\sum_{i=1}^{N}w_i h_i \quad (5)$$

$$w_i=\frac{\exp(\alpha_i)}{\sum\exp(\alpha_j)} \quad (6)$$

其中，h为BERT模型中间层输出；w为每一层的权重。

本研究将基于注意力的BERT层融合模型命名为BERT-ALA（Attention Based Layer Aggregation for BERT），后面的试验统一用这个名称，此机制可以应用在任意基于BERT的模型中。将

BERT-ALA应用在BERT+BiLSTM+CRF中，得到BERT-ALA+BiLSTM+CRF，主要结构如图3所示。BERT模型不同层的输出通过一组可以学习的权重参数加权得到最后的特征表示，再送入后续的BiLSTM以及CRF里面进行实体识别。

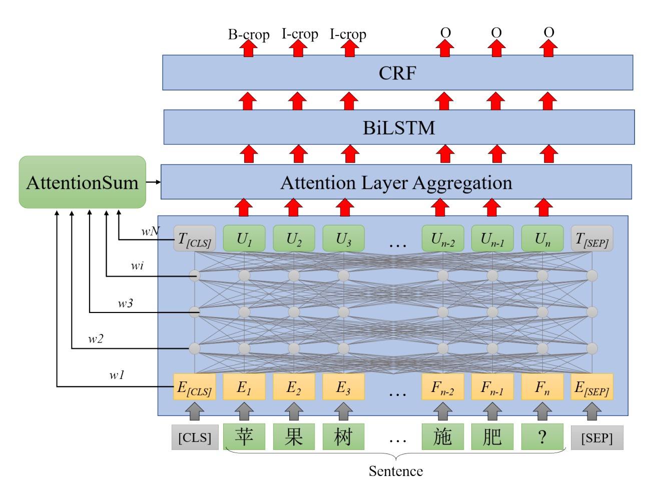


图3 BERT-ALA+BiLSTM+CRF架构图

Fig. 3 Architecture of BERT-ALA+BiLSTM+CRF

2.5 模型蒸馏

模型蒸馏[14]就是将训练好的复杂模型推广能力“知识”迁移到一个结构更为简单的网络中，或者通过简单的网络去学习复杂模型中的“知识”。前面提出了基于BERT的几种模型，但是BERT因为参数量大导致推理耗时高，很难满足实时推理需求。因此，本研究用BiLSTM+CRF作为学生模型，蒸馏前面提出的教师模型。相对于传统模型蒸馏只是蒸馏最后一层的输出而言，本研究还蒸馏了教师模型中间的BiLSTM层。蒸馏的损失函数一共分为3项，目标函数表达如下。

$$
\begin{aligned}
loss = {} & \alpha_1 MSEloss\left(h_{BiLSTM}(T), h_{BiLSTM}(S)\right) + \\
& \alpha_2 CEloss\left(h_{CRF}(T), h_{CRF}(S)\right) + \\
& \alpha_3 CRFloss\left(y_{true}, h_{CRF}(S)\right)
\end{aligned} \tag{7}
$$

其中，S表示学生模型；T表示教师模型；$h_{layer}(model)$表示model的layer层（BiLSTM层，CRF层）输出。因此，蒸馏损失的3项分别表示为：（1）学生模型BiLSTM层输出拟合教师模型BiLSTM层的输出，拟合损失是平均平方误差MSE；（2）学生模型CRF层输出的概率分布，与教师模型CRF层输出的概率分布求交叉熵；（3）原来的CRF损失[15]。其中，由CRF层输出概率与真实的实体识别标签计算得到。

2.6 模型推理

在推理阶段，接受用户端文本输入后，包含三个阶段的流程。

（1）通过词典匹配得到句子里面的农业类型实体S1。

（2）通过学生模型预测得到句子里面的农业实体S2。

（3）模型和词典得到的标注结果用求并集的方法聚合，返回给用户端；对于在S2而不在S1中的实体，是词典中还不存在的，返回人工专家复查，得到新词加入词典，以提高词典的覆盖率。

3 试验验证与分析

3.1 评价指标

试验指标采用精确匹配模式，被实体识别模型识别出来的称为mention，mention和ground truth里面的实体都表示为（start，end，type）的形式，start和end表示mention或者entity的边界，type表示类型。对于实体识别领域来说，TP、FP和FN的定义如下。

（1）True Positive（TP）：农业实体识别模型识别出来的mention，与ground truth里面的实体能对应上；

（2）False Positive（FP）：农业实体识别模型识别出来的mention，与ground truth里面的实体不能对应上，这里也包含了边界识别正确，但是类型识别错误的情况；

（3）False Negative（FN）：ground truth里面存在的entity，没有被农业实体识别模型识别出来。

根据上面定义的TP、FP和FN可以计算Precision、Recall和$F1$分数值，分别表示准确率、召回率以及$F1$分数值如下。

$$Precison = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

实体包含多种类型，不同类型的实体分别计算实体识别的$F1$，然后计算整体的$F1$，整体$F1$采用*macro-F*1 [26] 的方式计算，是各个类别$F1$的平均，公式如下所示。

$$macroF1 = \frac{\sum_{i=1}^{c} F1_i}{2} \tag{11}$$

3.2 试验设计

本研究选取了农业和医学两个领域，总共三个数据集来进行实验验证。选取医学领域的原因是因为医学领域与农业领域一样，都属于特定领域，另外，医学领域实体识别相对农业领域数据研究的比较多，容易获取开源的实验识别标注数据。第一个数据集是本研究构建的数据集，后面两个数据集是公开的数据集。

数据一：从互动百科获取的农业领域文本，按照句子进行切分。采用基于远程监督的方式构建训练集，验证集由人工标注。其中包含作物实体4662个，疾病实体695个。训练集和测试集的比例是8:2，训练集有10,277条数据，测试集有2532条数据。数据集已经在数据建模和数据分析竞赛平台kaggle上开源（https://www.kaggle.com/supportvectordevin/agriculture-pedia.）。

数据二：来源于讯飞开放平台的“农业问答数据处理挑战赛”里面的实体识别任务（http://challenge.xfyun.cn/topic/info?type=agriculture.）[25]，标注出农作物、病虫害和农药的命名实体标签。数据集包含病虫害实体100,660个，农药实体250,740，作物实体5796个。训练集包含15,624个样本，测试集包含3906个样本。

数据三：医学领域数据，来源于ccks 2017的task 2，面向电子病历的命名实体识别（Clinical Named Entity Recognition，CNER）（https://github.com/zjy-ucas/ChineseNER）。即对于给定的一组电子病历文档（纯文本文件），任务的目标是识别并抽取出与医学临床相关的实体名字。数据集包含症状和体征实体12,821个、检查和检验实体17,655个、疾病和诊断实体4560个、治疗实体4940个、身体部位实体17,556个。训练集包含10,787个样本，测试集包含2697个样本。

模型超参数方面，LSTM+CRF的词向量采用fast text Chinese word embedding [27]，LSTM隐含层数量是128。训练方面，采用Adam优化器 [28]，BERT层学习速率为10^{-5}，其他层为10^{-3}，batch size是32，每个batch采用batch内部最长的句子做padding，以减少内存消耗，但是最长截断长度设置为64。

3.3 基线模型对比验证

在三个数据集上，测试了三种基线模型的*macro-F*1，结果如表1所示。

表1 三种基线模型*macro-F*1对比

Table 1 Comparison of *macro-F*1 with three baseline models

模型	数据一	数据二	数据三
BiLSTM+CRF	0.8420	0.8027	0.8872
BERT+CRF	0.9195	0.9366	0.9036
BERT+BiLSTM+CRF	0.9266	0.9402	0.9105

分析验证结果得到三个结论。

（1）关于数据一的结果表明，远程监督的训练集训练的模型在人工标注的测试集上表现良好，证明了通过远程监督构建数据集的有效性。

（2）引入大规模预训练模型BERT相对于BiLSTM能显著提高模型的表现；相对于数据一的*macro-F*1提高7.75%，数据二的*marco-F*1提高13.39%，数据三是医疗实体识别数据，提升相对小一些，为1.64%。

（3）在BERT后面加入BiLSTM，能在一定程度上缓解BERT相对位置捕获不强的缺陷，在

数据一上，*macro-F*1 相对于 BERT+CRF 提高了 0.71%；在数据二上，提高了 0.36%；在数据三上，提高了 0.69%。

3.4 基于注意力的BERT层融合机制有效性验证

针对 BERT+CRF 和 BERT+BiLSTM+CRF 两个 BERT 系模型，分别用层融合机制改进 BERT，验证结果是否对实体识别结果有提高。结果如表2所示。

表2 层融合机制有效性验证结果

Table 2 Validation of layer aggregation mechanism

模型	数据一	数据二	数据三
BERT+CRF	0.9195	0.9366	0.9036
BERT-ALA+CRF	0.9293	0.9444	0.9153
BERT+BiLSTM+CRF	0.9266	0.9402	0.9105
BERT-ALA+BiLSTM+CRF	0.9360	0.9549	0.9237

验证结果表明，基于注意力的层融合机制在三个数据集上都能提高实体识别的效果。说明层融合机制在实体识别领域具有一定普适性。BERT-ALA+CRF 和 BERT-ALA+BiLSTM+CRF 相对于基准模型分别有大约 1% 的 *macro-F*1 的提高。

BERT-ALA+BiLSTM+CRF 在所有模型里面的效果最好，所以被选择为教师模型，指导蒸馏部分的学生模型学习。本研究主要是将 BERT-ALA+BiLSTM+CRF 应用在农业实体识别领域。

3.5 模型蒸馏效果验证

通过模型蒸馏的方法得到的教师模型是 BERT-ALA+BiLSTM+CRF，学生模型是 BiLSTM+CRF。与教师模型相比，学生模型的时间和空间复杂度都有改善。本研究用预测 1000 个样本的平均耗时表示模型的预测耗时，用于比较学生模型时间复杂度的改善；模型大小用模型占据的存储空间表示，用于验证学生模型空间复杂度的提高。由于这两个指标与数据无关，因此本研究在 3 个数据进行实验后取平均值。结果表明，蒸馏后的学生模型相对于教师模型每预测千个样本的耗时减少了 33%，模型大小减少了 98%，时间复杂度和空间复杂度都有了很大的改善，更加适用于线上预测场景。

本研究测试了蒸馏得到学生模型相对于用标注数据训练的同等模型的效果提高，*maro-F*1 指标的对比结果如表 3 所示。

表3 学生模型与教师模型 *macro-F*1 对比

Table 3 Comparison of *macro-F*1 with teacher model and student model

模型	数据一	数据二	数据三
BiLSTM+CRF	0.8420	0.8027	0.8872
Teacher Model	0.9360	0.9549	0.9237
Student Model	0.8730	0.8436	0.9154

验证结果表明，采用模型蒸馏的训练方法，相对于训练数据训练的同等模型，学生模型学到了更多的暗知识。蒸馏得到的学生模型在数据一上，*macro-F*1 提高了 3.1%。在数据二上，提高了 4.09%，在数据三上，提高了 2.82%。

3.6 学生模型效果展示

本研究主要应用场景是农业实体识别，因此以番茄为例，选取了几个番茄的百问百答[27]问句以及回答，验证最终线上蒸馏的学生模型效果，句子及其识别的结果如下。

提问 1：番茄病毒病症状及防治方法是什么？

识别结果：{'mention':'番茄病毒病', 'type':'disease', 'offset':0}

提问 2：番茄筋腐病是怎样产生的，如何防止？

识别结果：{'mention':'番茄筋腐病', 'type':'disease', 'offset':0}

提问 3：症状：番茄细菌性斑疹病主要危害叶、茎、花、叶柄和果实。

识别结果：{'mention':'番茄细菌性斑疹病', 'type':'disease', 'offset':3}

上述提问 1、2 和 3 的实体都能完整识别出来。其中，提问 2 和 3 中的实体“番茄筋腐病”

和“番茄细菌性斑疹病”都没有出现在词典中，也就是不存在于标注数据中，但是模型能识别成功，验证了模型具有良好的泛化性能。

4 结 论

本研究提出用远程监督构建农业实体识别数据，标注数据存在漏标注的问题。基于漏标注的句子远比标注正确的句子少的假设，解决的思路是用弱标注的数据训练一个初级版本的实体识别模型，再用实体识别模型选择训练集里面一些置信度低的结果，返回来进行校正，最后用校正后的数据对基础版本模型进行微调。

（1）主要研究了农业领域的实体识别问题。针对农业领域缺乏实体识别标注数据的问题，提出爬取互联网开源数据库“互动百科”构建农业知识图谱，远程监督实现实体识别数据弱标注的方案。

（2）针对过往研究使用的模型识别效果不佳、依赖手工特征的问题，结合农业实体识别的特点，提出了基于注意力层融合机制的 BERT-ALA+BiLSTM+CRF 模型，在 3 个数据集上都取得了最优的效果，验证了层融合机制的有效性。本研究的目的主要是将这个模型应用在农业实体识别领域。

（3）针对基于 BERT 的模型预测耗时长的问题，用 BiLSTM+CRF 模型作为学生模型提出的蒸馏 BERT-ALA+BiLSTM+CRF 模型，大大降低了线上模型的时间复杂度和空间复杂度，使得训练后的模型在移动端应用成为可能。

本研究提出的实体识别方法在解决农业领域实体识别问题方面实现了农业智能化究方法，还可以拓展应用到其他标注数据缺失的垂直领域实体识别场景，如医学、教育、军事等。

致 谢

感谢中国农业科学院蔬菜与花卉研究所贺超兴研究员为本研究提供意见。

参考文献：

[1] COWIE J, LEHNERT W. Information extraction[J]. Communications of the ACM, 1996, 39(1): 80-91.

[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics, 2016: ID N16-1030.

[3] 李贯峰, 张鹏. 一个基于农业本体的 Web 知识抽取模型[J]. 江苏农业科学, 2018, 46(4): 201-205.
LI G, ZHANG P. A web knowledge extraction model based on agricultural ontology[J]. Jiangsu Agricultural Sciences, 2018, 46 (4): 201-205.

[4] 王春雨, 王芳. 基于条件随机场的农业命名实体识别研究[J]. 河北农业大学学报, 2014, 37(1): 132-135.
WANG C, WANG F. Research on agricultural named entity recognition based on conditional random field[J]. Journal of Hebei Agricultural University, 2014, 37 (1): 132-135.

[5] TSENG H, CHANG P-C, ANDREW G, et al. A conditional random field word segmenter for sighan bakeoff 2005[C]// Proceedings of the fourth SIGHAN workshop on Chinese language Processing. San Diego, USA: Association for Computational Linguistics, 2005.

[6] MALARKODI C, LEX E, DEVI S L J. Named entity recognition for the agricultural domain[J]. Research in Computing Science, 2016, 117(1): 121-132.

[7] 刘晓俊. 面向农业领域的命名实体识别研究[D]. 合肥: 安徽农业大学, 2019.
LIU X. Research on named entity recognition for agriculture[D]. Hefei: Anhui Agricultural University, 2019.

[8] BISWAS P, SHARAN A, VERMA S. Named entity recognition for agriculture domain using word net[J]. IInternational Journal of Computer & Mathematical Sciences 2016, 5(10): 29-36.

[9] MILLER G A. WordNet: An electronic lexical database[M]. Massachusetts: MIT press, 1998.

[10] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge Data Engineering, 2020 (99): 1.

[11] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]// Proceedings of the Joint Conference of the 47th Annu-

al Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. San Diego, USA: Association for Computational Linguistics, 2009: 1003-1011.

[12] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]// Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1753-1762.

[13] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics, 2018.

[14] POLINO A, PASCANU R, ALISTARH D. Model compression via distillation and quantization[EB/OL]. 2018. arXiv:1802.05668.

[15] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. 2015. arXiv: 1508.01991.

[16] ZHOU Z. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44-53.

[17] 米嘉. 大规模中文文本检索中的高性能索引研究[D]. 北京: 中国科学院, 2005.

MI J. Research on high performance index in large scale Chinese text retrieval[D]. Beijing: Chinese Academy of Sciences, 2005.

[18] LUO L, YANG Z, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34 (8): 1381-1388.

[19] SOUZA F, NOGUEIRA R, LOTUFO R. Portuguese named entity recognition using BERT-CRF[EB/OL]. 2019. arXiv:1909.10649.

[20] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28 (10): 2222-2232.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, US: Carran Associates Inc., 2017: 6000-6010.

[22] DONG C, ZHANG J, ZONG C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]//International Conference on Computer Processing of Oriental Languages National CCF Conference on Natural Language Processing and Chinese Computing. Berlin, German: Springer, 2016: 239-250.

[23] YAN H, DENG B, LI X, et al. Tener: Adapting transformer encoder for name entity recognition[EB/OL]. 2019. arXiv:1911.04474.

[24] JIANG S, ZHAO S, HOU K, et al. A BERT-BiLSTM-CRF model for chinese electronic medical records named entity recognition[C]// 2019 12th International Conference on Intelligent Computation Technology and Automation. Piscataway, New York, USA: IEEE, 2019: 166-169.

[25] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. San Diego, USA: Association for Computational Linguistics, 2019.

[26] OPITZ J, BURST S. Macro F1 and Macro F1[EB/OL]. 2019. arXiv:1911.03347.

[27] GRAVE E, BOJANOWSKI P, GUPTA P, et al. Learning word vectors for 157 languages[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA) , 2018.

[28] KINGMA D P, BA J J A P A. Adam: A method for stochastic optimization[EB/OL]// 3rd International Conference on Learning Representations. Ithaca, NY: arXiv. org. 2015: 13.

Agricultural Named Entity Recognition Based on Semantic Aggregation and Model Distillation

LI Liangde[1,2], WANG Xiujuan[1,3], KANG Mengzhen[1,2*], HUA Jing[1,4], FAN Menghan[1,2]

(1. *The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190*, China*;2. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100049*, China*;3. *Beijing Engineering Research Center of Intelligent Systems and Technology, Beijing* 100190*, China*;4. *Qingdao Smart AgriTech. ,Ltd, Qingdao* 266000*, China*)

Abstract: With the development of smart agriculture, automatic question and answer (Q&A) of agricultural knowledge is needed to improve the efficiency of agricultural information acquisition. Agriculture named entity recognition plays a key role in automatic Q&A system, which helps obtaining information, understanding agriculture questions and providing answer from the knowledge graph. Due to the scarcity of labeled ANE data, some existing open agricultural entity recognition models rely on manual features, can reduce the accuracy of entity recognition. In this work, an approach of model distillation was proposed to recognize agricultural named entity data. Firstly, massive agriculture data were leveraged from Internet, an agriculture knowledge graph (AgriKG) was constructed. To overcome the scarcity of labeled named agricultural entity data, weakly named entity recognition label on agricultural texts crawled from the Internet was built with the help of AgriKG. The approach was derived from distant supervision, which was used to solve the scarcity of labeled relation extraction data. Considering the lack of labeled data, pretraining language model was introduced, which is fine tuned with existing labeled data. Secondly, large scale pretraining language model, BERT was used for agriculture named entity recognition and provided a pretty well initial parameters containing a lot of basic language knowledge. Considering that the task of agriculture named entity recognition relied heavily on low-end semantic features but slightly on high-end semantic features, an Attention-based Layer Aggregation mechanism for BERT(BERT-ALA) was designed in this research. The aim of BERT-ALA was to adaptively aggregate the output of multiple hidden layers of BERT. Based on BERT-ALA model, Bidirectional LSTM (BiLSTM) and conditional random field (CRF) were coupled to further improve the recognition precision, giving a BERT-ALA+BiLSTM+CRF model. Bi-LSTM improved BERT's insufficient learning ability of the relative position feature, while conditional random field models the dependencies of entity recognition label. Thirdly, since BERT-ALA+BiLSTM+CRF model was difficult to serve online because of the extremely high time and space complexity, BiLSTM+CRF model was used as student model to distill BERT-ALA+BiLSTM+CRF model. It fitted the BERT-ALA+BiLSTM+CRF model's output of BiLSTM layer and CRF layer. The experiment on the database constructed in the research, as well as two open datasets showed that (1) the *macro-F*1 of the BERT-ALA + BiLSTM + CRF model was improved by 1% compared to the baseline model BERT + BiLSTM + CRF, and (2) compared with the model trained on the original data, the *macro-F*1 of the distilled student model BiLSTM + CRF was increased by an average of 3.3%, the prediction time was reduced by 33%, and the storage space was reduced by 98%. The experimental results verify the effectiveness of the BERT-ALA and knowledge distillation in agricultural entity recognition.

Key words: distant supervision; agriculture knowledge graph; agriculture Q&A system; named entity recognition; knowledge distillation;deep learning; BERT; Bi-LSTM